ABSTRACT OF THE DISCLOSURE

The present invention relates to a system and method for classifying documents in order to select the most desirable documents of a group. Because quality is very difficult to distinguish by anyone other than a human being, this invention provides a system and method that will create a profile of what constitutes quality, then utilize this profile to allow a user to retrieve information that is desirable. A client is provided with items of data selected according to estimates computed using a profile of certain high-level criteria such as quality, interestingness, appropriateness, timeliness, humor, style of language, obscenity, sentiment, and any combinations thereof. These estimates are computed from low-level criteria such as length, vocabulary, fraction of words spelled correctly, title, author, reading grade level, average length of sentences, average length of words, usage of punctuation, usage of grammar, formatting, capitalization, source, display tags and any combinations thereof. The profile is learned automatically from labeled training examples.

This system also relates to a method of obtaining and automatically associating a value to an item of data by obtaining items, obtaining labels for some items, selecting items of data with certain labels to form training sets, learning a profile using the training sets, and associating a value to another item of data using said profile. As such, the program is capable of learning to measure which items are of high quality and is capable of delivering only those items of data which would be of interest to a client.